

Towards Demystifying Subliminal Persuasiveness: Using XAI-Techniques to Highlight Persuasive Markers of Public Speeches

Klaus Weber¹[0000–0001–6661–8602], Lukas Tinnes², Tobias Huber¹[0000–0002–5010–4006], Alexander Heimerl¹, Marc-Leon Reinecker³, Eva Pohlen², and Elisabeth André¹[0000–0002–2367–162X]

¹ Augsburg University, Augsburg, Germany
{weber, huber, heimerl, andre}@hcm-lab.de

² Augsburg University, Augsburg, Germany
{name}. {surname}@student.uni-augsburg.de

³ University of Applied Sciences Augsburg, Augsburg, Germany
marc-leon.reinecker@hs-augsburg.de

Abstract. The literature provides evidence for the importance of non-verbal cues when it comes to persuading other people and developing persuasive robots. Mostly, people use these non-verbal cues subconsciously and, more importantly, are not aware of the subliminal impact of them. To raise awareness of subliminal persuasion and to explore a way for investigating persuasive cues for the development of persuasive robots and agents, we have analyzed videos of political public speeches and trained a neural network capable of predicting the degree of perceived convinciness based on visual input only. We then created visualizations of the predictions by making use of the explainable artificial intelligence methods Grad-CAM and layer-wise relevance propagation that highlight the most relevant image sections and markers. Our results show that the neural network learned to focus on the person, more specifically their posture and contours, as well as on their hands and face. These results are in line with existing literature and, thus, show the practical potential of our approach.

Keywords: Subliminal Persuasion · Persuasive Markers · XAI

1 Introduction

In the process of changing opinions or attitudes, people use far more than logical and rational aspects. There is evidence from the literature that the persuasive power of arguments largely depends on appropriate body language. Consequently, if arguments that are content-wise identical are presented differently, i.e. with different non-verbal behaviors, the persuasive power of an argument can be different.

There is significant evidence from the literature that body language and the type of gestures used influence how a person is perceived, and several studies

showed that body language and verbal aspects significantly influence perceived persuasiveness [5, 9, 14]. These body-language-based cues, however, are often unconsciously observed by people, and it seems that people are not aware of this kind of subliminal persuasion.

Understanding these cues bears two advantages: 1) It can help people behave differently, i.e., more persuasive, in debates, speeches or job interviews, and 2) a deeper understanding of these persuasive cues can help researchers develop persuasive robots and agents in human-robot interactions more easily [9, 14].

In this paper, we explore an approach employing explainable artificial intelligence techniques to make persuasive cues visible to demonstrate the importance of the persuasive power of body-language-based argumentation and to investigate a different approach to developing persuasive agents and robots.

First, we trained a model to predict perceived convincingness based on an annotated political public speech using a convolutional neural network utilizing the visual (image) channel only (i.e., without the audio channel). We then employed explainable artificial intelligence (XAI) visualization techniques to uncover what parts of the image were the most relevant ones for predicting the degree of perceived convincingness.

Our post-hoc analysis reveals that our neural network has learned to focus on the person speaking and (mostly) ignore the background of the image. The observations of our visualizations indicate that the network primarily localizes hand and face positions on the image, which demonstrates, in line with existing literature, the importance of subliminal persuasive cues.

The structure of the paper is as follows. Section 2 gives an overview of persuasion theory and XAI visualization techniques, Section 3 describes the overall approach, including the data annotation process and the architecture of the trained model. Section 4 highlights what the network has learned employing Grad-CAM and Layer-wise Relevance Propagation (LRP). Finally, Section 5 concludes with a brief discussion of results, limitations of our approach, and future work.

2 Related Work

Related work of this research can be divided into two parts: (1) The effect of non-verbal cues in persuasive messages and (2) Explainable Artificial Intelligence.

2.1 The Effect of Non-Verbal Cues in Persuasive Messages

The theory of persuasion goes back to Aristotle. He identified three means of persuasion, namely logos, pathos, and ethos. Logos defines the logical and rational aspects, i.e., the content of the argument, pathos the emotional engagement between the speaker and the listener, while ethos describes the personality of the speaker, their character, and how the speaker is perceived by the audience [20].

According to psychological models, there are two cognitive routes (*central* and *peripheral*), through which a persuasive message can be processed. Petty and Cacioppo [26] developed the Elaboration Likelihood Model (ELM) describing

the influence of information processing on the result of a persuasive message depending on the listeners’ “*need for cognition*” (NFC). If the listener’s NFC is low, then a message is more likely processed via the *peripheral route* otherwise *central processing* takes place. Chaiken et al. [8] extended this model (Heuristic-Systematic Model – HSM) claiming that people do not process information in isolation via one of the two routes. Instead, peripheral processing always takes place, to which central processing is added when an elaboration threshold is reached (depending on the listener’s *need for cognition*).

Consequently, researchers have investigated the effect of non-verbal cues on the perceived persuasiveness. DeSteno et al. [10] showed that persuasive messages are more successful if they are framed with emotional overtones that correspond to the emotional state of the recipient. Wang et al. [33] showed that perceived persuasiveness of emotions depends on the level of power of the speaker and the listener. Further, Van Kleef et al. [18, 32] showed that people use the source’s emotions as information channel when they form their attitudes.

In addition to that, researchers have investigated the effect of gestures and gaze. Maricchiolo et al. [22] investigated the effect of hand gestures concerning the speaker’s perceived persuasiveness revealing that hand gestures affect the evaluation of a message’s persuasiveness, the speaker’s style effectiveness, and their composure and competence. Poggi et al. [27] further investigated the use of gestures and gaze in political discourse concerning their persuasive import.

In short, there is a lot of evidence that persuasiveness largely depends on body-language-based argumentation and persuasive cues. Thus, by taking away the audio channel, a neural network should be able to learn these cues to predict perceived persuasiveness successfully. Hence, in this paper, we investigate 1) whether or not a neural network can “*understand*” and learn these subliminal cues and 2) whether or not the network learns to focus on the sections containing these subliminal cues instead of focusing on the image as a whole.

2.2 Explainable Artificial Intelligence

Since artificial intelligent systems are becoming more and more complex, there is an increasing need to increase the explainability of these systems. Understanding how a system works is crucial for working with and building trust in artificial, intelligent systems.

XAI is especially important when the system is inferring personality traits of humans, such as persuasiveness, which is a highly subjective task that might include biases. For this reason, earlier works used XAI on several subjective tasks. Escalante et al. [12], for example, developed a challenge to test different explainable systems that are used for first impression analysis in the context of job applications. Weitz et al. [34] investigated different XAI methods on facial pain and emotion recognition models. However, to the best of our knowledge, this is the first work on explainable systems that predict the degree persuasiveness of humans. In the context of persuasion and XAI, recent work mainly investigated explainable recommendation systems persuading humans [11, 36].

XAI is often split into several subcategories. In this work, we do not, for example, deal with the development of more interpretable model architectures. Instead, we focus on *post hoc* explanations that are created after the model was trained [23]. Furthermore, we focus on local explanations that analyze single predictions of a system instead of global explanations that try to shed light on the general behavior of a system. For neural networks, the most common local post-hoc explanation method is the generation of saliency maps [1]. Saliency maps are heat-maps that highlight areas of the input that were relevant for the decision of a system in a certain way.

One of the first kinds of saliency maps were based on the gradient. Simonyan et al. [30] used backpropagation to calculate the gradient with respect to each input unit to measure how much a small change in this input affects the prediction. Selvaraju et al. [29] made this approach more class discriminatory by stopping the backpropagation after the fully connected layers and using the gradient with respect to the output of the last convolutional layer.

A different kind of saliency map estimates how much each input attributed to the final decision of a neural network. Lapushkin et al. [6, 21] introduced layer-wise relevance propagation (LRP) that assigns a relevance value to each neuron in a neural network, measuring how relevant this neuron was for a particular prediction. For this assignment, they defined different rules, all of which are based on the intermediate outputs of the neural network during the forward pass. One of those rules introduced by Huber et al. [15] tries to create more selective saliency maps by only propagating the relevance to the neuron with the highest activation in the preceding layer. Montavon et al. [24] put the LRP concept into the theoretical framework of the Taylor decomposition.

Another take on saliency maps comes with occlusion or perturbation based visualizations. Zeiler et al. [35] zero out windows inside the input and measure how much the prediction changes. The more the output changes, the more relevant was this window for this particular prediction. Greydanus et al. [13] uses a similar approach but perturbs the windows with noise to see how much the introduced uncertainty affects the prediction. The LIME framework from [28] first separates the input picture into super-pixels by a segmentation algorithm. Afterwards, a more interpretable model is trained to estimate which super-pixels are the most relevant for a given decision. One of the advantages of those methods is that they are not dependant on the structure of the model, but this comes with the drawback of not being as precise as some model-specific methods.

Recently, Adebayo et al. [2] introduced a sanity check that showed that some gradient-based saliency maps were not analyzing the learned weights of a neural network. The original saliency maps from [30] and the Grad-CAM maps both passed the test. This year, Sixt et al. [31] tested different LRP variants more in depth. They concluded that most LRP variants lose a lot of information about the last fully connected layers of the network. Instead, they mainly analyze the convolutional layers at the beginning of the network. Therefore we chose a combination of class discriminatory Grad-CAM saliency maps and fine granular

LRP saliency maps to get a good understanding of the end and the beginning parts of our model respectively.

3 Data Annotations and Model

In this Section, we describe the data annotation process and the model architecture, including the training process of the neural network in detail.

3.1 Corpus and Annotation Process

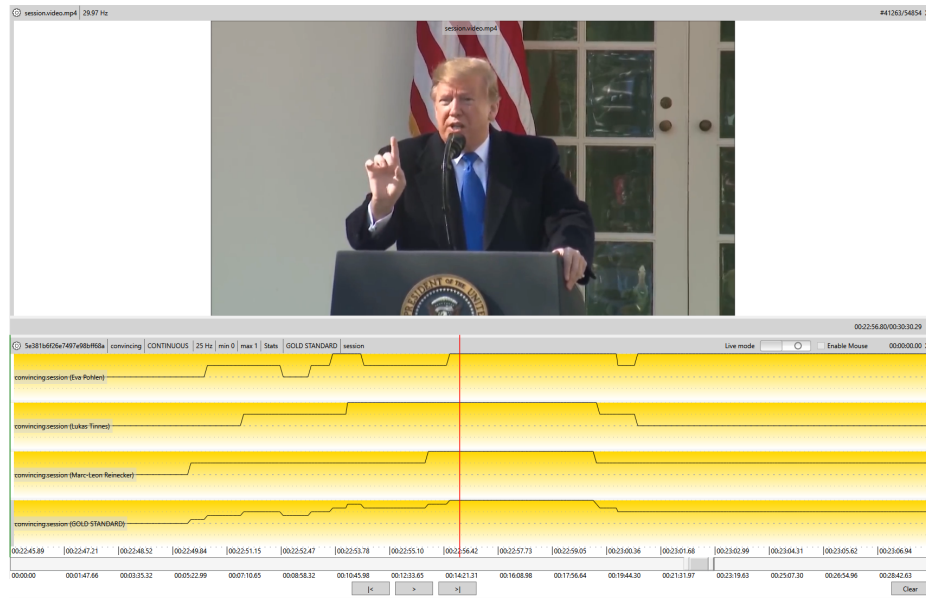


Fig. 1. Screenshot of the NOVA tool depicting the video at the top and four annotation streams below (3 annotators + merged gold standard for the training process).

The training corpus consists of a public speech by Donald J. Trump, which was held in 2019 with an approximate length of 50 minutes.¹ The data were annotated using NOVA [7], an annotation tool for annotating and analyzing behavior in social interactions. The NOVA user interface has been designed with a particular focus on the annotation of continuous recordings involving multiple modalities and subjects. It supports several techniques from the latest developments from contemporary research fields such as Cooperative Machine Learning and XAI to enhance the standard annotation process. We had the corpus continuously annotated by three experienced labelers with a sample rate of 25Hz.

¹ <https://www.youtube.com/watch?v=DU6BnuyjJqI>

They were asked to rate how convincing the speaker appeared distinguishing between five different levels (ranging from *not convincing at all* to *very convincing*). The annotators achieved an inter-rater agreement of 0.77 (Cronbach’s α), which seems sufficient for our purpose considering the high subjectivity of perceived persuasiveness [16, 25]. The final annotations have been merged (see Figure 1) to obtain a gold standard annotation stream with more than 50,000 sample images. Due to the nature of the video, the lowest two classes were barely annotated.

3.2 Model Architecture and Training

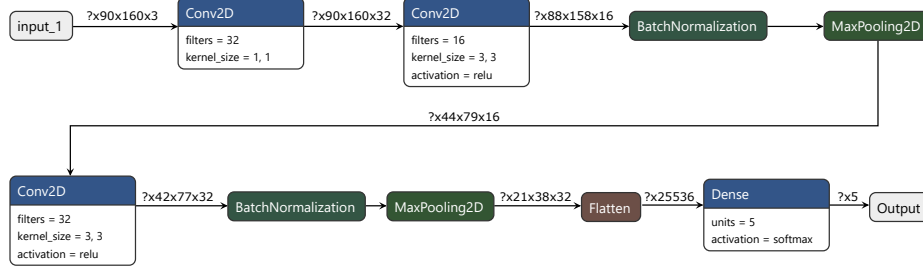


Fig. 2. An illustration of the network architecture. The network consists of three convolutions, which learn to focus on body parts important for predicting convincings. The first layer expands the 3-channel RGB to 32 channel before being fed into the last two convolutions layers, after each of which batch normalization and max-pooling are applied. The network outputs a 5-vector estimating the probability of each class.

Figure 2 sketches the architecture of our employed convolutional neural network consisting of three subsequent convolutional layers. The last two layers are followed by batch normalization and max-pooling layers. The output of the last convolutional layer is flattened and then fed into a five-way softmax function to get the predictions of all five classes.

We first extracted the video frames with a sample rate of 25Hz and down-sampled them to 160x90 RGB-Images. The first convolutional layer expands the RGB-channel of the input image to 32 channels. The idea behind this is that we allow the network to define colors for different pixel combinations similar to how humans see, for example, a combination of yellow and blue as green. The network outputs a five-dimensional vector describing the probability of each class. A ReLU activation is used in each layer apart from the output layer, in which a softmax function is applied. As optimizer we use Adamax ($\beta_1 = 9$, $\beta_2 = 0.999$) [17].

To tackle overfitting, we use batch normalization as well as L2-regularization. Batch normalization is applied after the second and third convolutional layer,

followed by pooling layers. L2-regularization (regularization factor 0.01), on the other hand, is applied to each convolutional layer in the network.

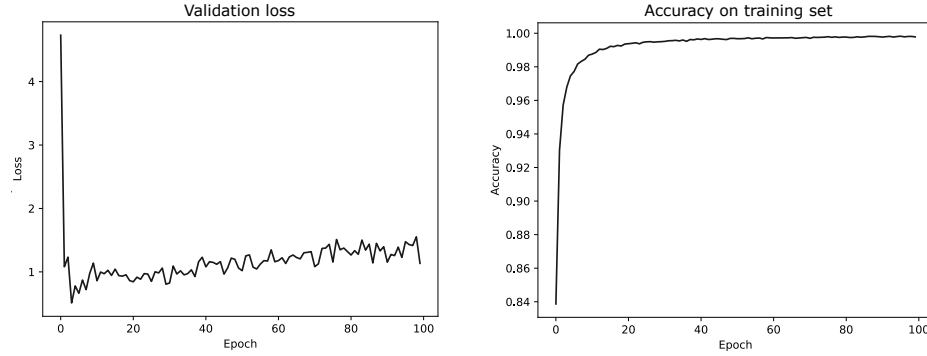


Fig. 3. Validation loss and training accuracy over all epochs.

The model was trained for 100 epochs using a batch size of 32 and with the dataset split into training and validation data by a ratio of 4:1.

Figure 3 summarizes the learning process showing that the neural network was able to predict classes reliably after only 20 epochs with an accuracy of $> 98\%$ on the training set. Since the validation loss shows slight overfitting after 20 epochs, the network explored in this work was only trained for 20 epoch.

To validate the performance of the network, we computed the confusion matrix on the training data set as visualized in Figure 4.

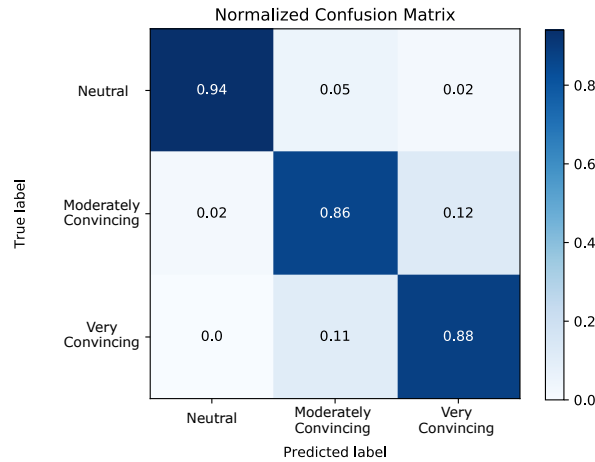


Fig. 4. Confusion matrix computed on the training data set to ensure that our network is sufficiently accurate on the learned samples.

Note that we have not trained a general predictor for persuasiveness as we only intend to explore what our network looks at when learning perceived persuasiveness. Therefore, we evaluated our model on the training data set only to ensure that our network is sufficiently accurate on the learned samples. Since the lowest two classes were not annotated at the current stage, they are not listed in the matrix.

We verified the performance of our model by computing the F1-scores indicating that our model performs very well on the learned samples (Table 1).

Measure	Class		
	Neutral	Moderately Convincing	Very Convincing
Precision	0.93	0.93	0.77
Recall	0.94	0.86	0.88
F1-Score	0.93	0.89	0.82

Table 1. Correlation between feedback and effectiveness.

4 Highlighting the Cues: Visualising the Network’s Eyes

Since we trained the network on images only, it seems that it was able to learn features that describe the perceived convincingness of a person. The interesting question is, which sections were the most relevant for making a (correct) prediction and if there are features that are in line with existing literature, i.e., did the network learn to focus on image excerpts that are evidenced indicators for perceived convincingness? To investigate this, we applied two different XAI techniques: (1) Grad-CAM and (2) Layer-wise Relevance Propagation.

4.1 Grad-CAM

To explain the predictions, we first analyzed the last layer of our network employing Grad-CAM [29] using keras-vis [19], a high-level toolkit for visualizing trained neural network models. For better visualizations, we created edge images of the input images and placed the network’s visualization maps over them.

Several example visualizations of different classes are depicted in Figure 5. They show that the network has learned to focus on the person, more specifically, their posture and contours. The background is mostly ignored and not relevant for the prediction (apart from a little background noise). More specifically, the network follows the hands and face of the speaker, which is in line with existing literature strengthening the validity of our approach since literature states that gestures, gaze, and hand movements are important indicators for perceived persuasiveness. It is worth noting that when predicting the *neutral* class, the network seems to look at every object on the image (unlike the other two classes where the network follows explicitly the person’s arms and hands of the person). This is probably since the network cannot find any convincing markers at all,

so every part of the image is observed. These visualizations inherently reveal the existence of a link between the visual channel and subliminal persuasion as well as the ability of neural networks to learn this connection demonstrating the importance of the persuasive power of non-verbal cues.

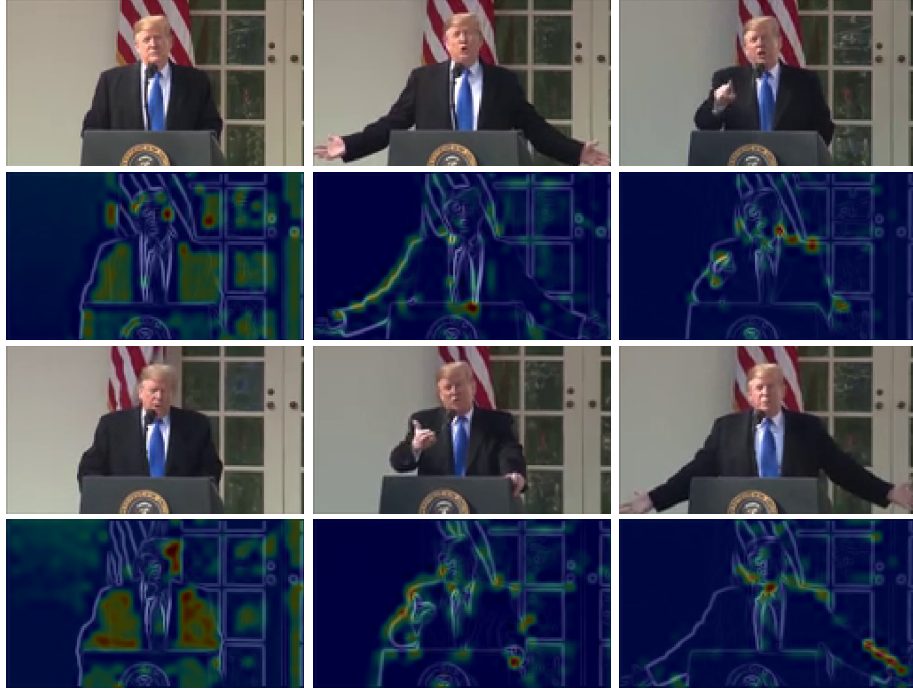


Fig. 5. Example visualization - (FLTR): Neutral - Moderately Convincing - Very Convincing. The visualization shows that the neural network has learned to focus on the posture, hands, and contours of the speaker to make its prediction. Due to the nature of our training data, the network hardly learned the person's features for barely annotated classes.

To examine the generalization of the network (despite being trained on one person only), we also tested the prediction on several images of other politicians, namely American senator Bernie Sanders², President of France Emmanuel Macron³ and Chancellor of Germany Angela Merkel⁴. The visualizations are depicted in Figure 6.

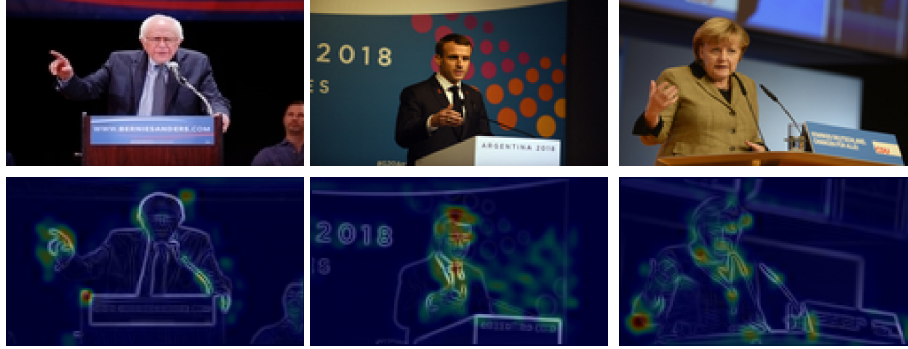


Fig. 6. Example visualizations of several other politicians with varying degrees of convincingness. (FLTR): Bernie Sanders (predicted class: *very convincing*) - Emmanuel Macron (predicted class: *very convincing*) - Angela Merkel (predicted class: *moderately convincing*).

Despite the speakers and the camera angle being different, the network still focuses on hands and the general face area. Taking a closer look at the picture of Emmanuel Macron reveals that the network seems to have learned to locate areas with skin-related colors to make its decision, even though the network does not always locate all image parts with skin-related color.

² Modification of 'Election 2016: Bernie Sanders NYC Fundraiser Draws Campaign Supporters Who Are 'Feelin' The Bern' by Michael Vadon: <https://flickr.com/people/80038275@N00/>, licensed under a Creative Commons License: <https://creativecommons.org/licenses/by-sa/2.0/>

³ Modification of 'Conferencia de Prensa - Presidente Emmanuel Macron - Día 2' by G20 Argentina: <https://www.flickr.com/photos/g20argentina/>, licensed under a Creative Commons License: <https://creativecommons.org/licenses/by/2.0/>

⁴ Modification of 'Rede der Bundeskanzlerin Angela Merkel zum Abschluss des CDU-Parteitage' by CDU/CSU Bundestagsfraktion, licensed under a Creative Commons License: <https://creativecommons.org/licenses/by-sa/3.0/deed.en>

4.2 Layer-wise Relevance Propagation

Next to Grad-CAM, we used LRP to analyze further the first convolutional layers of the network and what patterns they learned. LRP assigns a relevance value R_k to each neuron in a neural network. Let a_k be the activation of the k -th neuron during the forward pass and let w_{jk} be the weight that connects neuron j and neuron k . After the forward pass, the relevance propagation starts in the output layer. Here, the activation responsible for the prediction gets assigned its activation as relevance and every other neuron gets set to zero. That is

$$R_k = \begin{cases} a_k & \text{if } k = \operatorname{argmax}\{a_k\} \\ 0 & \text{if not.} \end{cases} \quad (1)$$

Beginning from there the relevance gets propagated to each preceding layer according to different rules (see Fig. 7). In our experiments we used the z^+ - or $\alpha 1\beta 0$ -rule:

$$R_j = \sum_k \frac{(a_j w_{jk})^+}{\sum_j (a_j w_{jk})^+} R_k, \quad (2)$$

where $(a_j w_{jk})^+$ is defined as $\max(a_j w_{jk}, 0)$.

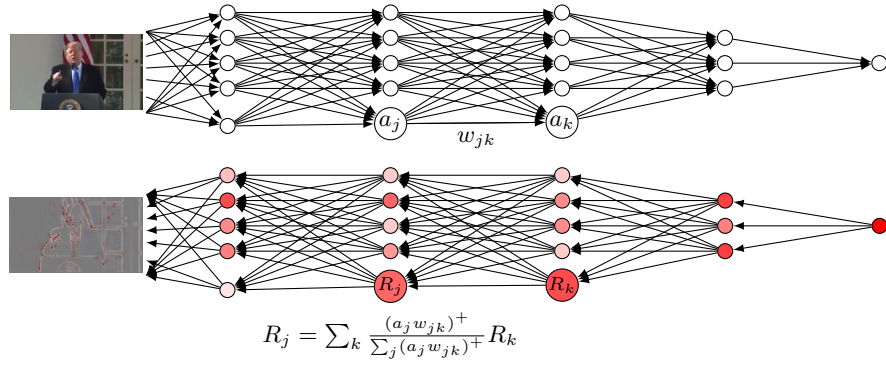


Fig. 7. Relevance propagation using the z^+ -Rule (Equation 2).

To create the LRP saliency maps for our model, we used iNNvestigate [3], a library that provides out-of-the-box implementations of many analysis methods, including LRP. Example visualizations can be seen in Figure 8. LRP visualizations show similar results as Grad-CAM. As before, we can see that the network seems to have learned the spatial features of the person, namely facial features, hand gestures, and the contour of the person. This again demonstrates the importance of subliminal persuasive cues in line with the literature and shows that neural networks are able to learn them.

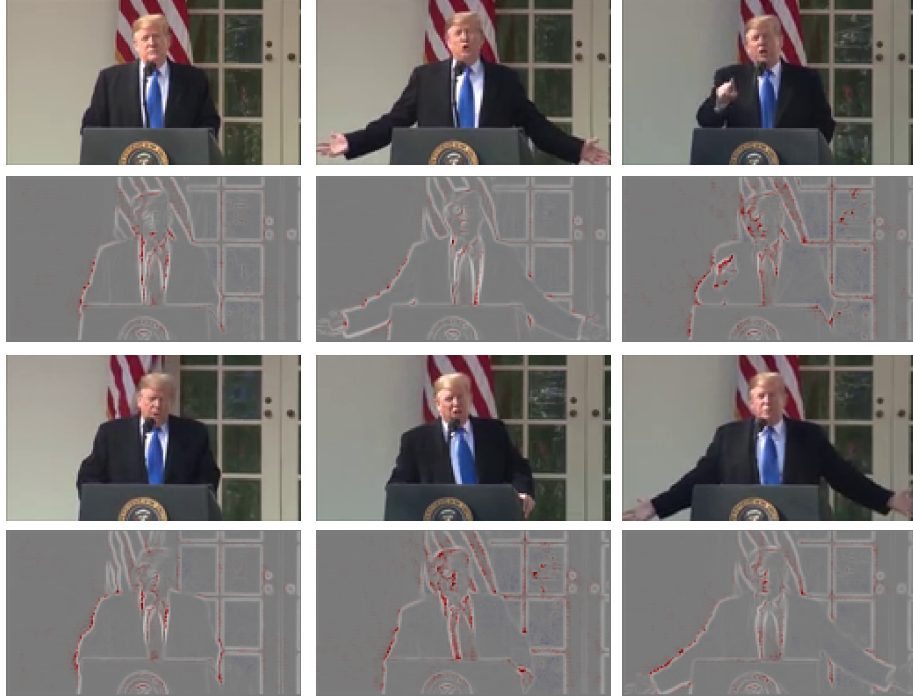


Fig. 8. Example LRP Visualizations (z+-rule) - (FLTR): Neutral - Moderately Convincing - Very Convincing.

5 Discussion and Limitations

In the beginning, we have argued that people are often persuaded by subliminal cues and that mostly they are not aware of them. To raise awareness of the existence of this subliminal persuasion, we have analyzed original political speeches and had annotators label them regarding their perceived convincingness by both listening and watching the video. We then trained a convolutional neural network on visual input only to predict the degree of convincingness and used XAI techniques, more specifically Grad-CAM and Layer-wise Relevance Propagation to highlight the most relevant sections. The results are fascinating, revealing that the network has not only learned to focus on the person and their contours but also the face and hands. The latter one is especially interesting as it shows, in line with existing literature, the importance of hand movements and, thus, demonstrates the importance of these subliminal persuasive cues. These results are, therefore, interesting for human-robot-interactions as they enable a different approach to investigating what makes humans persuasive and how to replicate these results in robots.

Apart from these preliminary results, our approach still faces some limitations that should not be neglected.

Limited Training Corpus. Our corpus consisted of only 50,000 samples of the same person; thus, it is unlikely that the network has learned a generalization for predicting the general degree of perceived persuasiveness, even though it also worked on some example images that the network has not seen before. We pointed out that we only tested the model on the training data set since the purpose of the model was to explore what parts of the image the network focuses on when learning perceived persuasiveness. In this regard, the results of the model training should be interpreted with some care, and it should not be considered a general predictor for perceived persuasiveness.

Even though our visualizations have shown that the network has learned to focus on hand and face positions, mainly by focusing on sections with skin-related colors. It is therefore questionable how well the current, trained model works on images with very light, skin-colored backgrounds. Since our network has also been trained on white skin color, the network would probably not work on people with other skin colors yet. Therefore, our data set needs several extensions, that are 1) adding data from different people with different skin colors and 2) adding data with different backgrounds to force the network to learn better generalization of convincing indicators.

No sequential Persuasive Indicators can be learned. The current approach uses a convolutional neural network for predicting the perceived persuasiveness based on a single input image only. However, there may be many persuasiveness indicators, such as the speed of hand movements which also influence perceived persuasiveness which cannot be learned with the current approach at present. Thus, in future work, we will further explore how we can highlight sequential types of persuasive markers using XAI techniques, such as LRP similar to Anders et al. [4].

Distribution of the Annotation Data and Annotation Process. Our annotated data consisted of only three classes: *neutral*, *moderately convincing*, and *very convincing*. Therefore, the network has not learned any characteristics yet about what *not convincing* people look like. Using only one video, this is expected, because from a common-sense perspective individuals may generally perceive another person as either more convincing or less convincing (exclusive-or). Also, the whole annotation process is subject to the annotator’s own opinion as persuasiveness, in general, is highly subjective. Therefore, it remains unclear, whether or not the annotators have annotated the perceived persuasiveness in general or just the intensity of the body language movement, which may also have an impact on the perceived persuasiveness. This limitation requires further analysis and will be addressed further in our future work. Also, we will explicitly include samples of the missing classes (i.e., different videos of other people) to obtain more detailed training results and to compare the markers of a *convincing* and *not convincing* appearance of people.

Nevertheless, our first results have shown the feasibility and practical potential of highlighting persuasive cues and indicators for persuasiveness employing explainable AI techniques.

6 Conclusion

In this paper, we explored an approach that highlights persuasive indicators of public speeches using explainable artificial intelligence techniques. There is a lot of evidence from the literature that bodily cues play an important role in persuading people. However, since people often seem not to be aware of the importance of body-language-based argumentation, we trained a convolutional neural network, which can predict perceived persuasiveness solely based on visual input. We then applied explainable AI techniques, namely Grad-CAM and Layer-wise Relevance Propagation in order to highlight relevant areas of the image that were used by the network for predicting the degree of persuasiveness to raise awareness of the stated importance of subliminal persuasive cues. Further we aim to explore an effective way for investigating persuasive cues for the development of persuasive agents and robots. Our results show that our network has learned to focus on the person, their contours, face, and hands proving that our network is able to look for parts on the image that are important indicators for a person’s persuasiveness according to existing literature. We have described the limitations of our approach in detail, especially concerning our used training data set, which only consisted of one speech of a single person. In our future work, we will address the limitations mentioned above and extend our corpus⁵ with additional speeches and look for suitable existing corpora to generalize our approach. We will then explore if our network can learn generalized as well as more fine-grained persuasive indicators, such as making a fist as well as sequential persuasive markers and if we can highlight such persuasive markers. Additionally, we will make use of other explainable AI techniques to get a deeper understanding of the impact of persuasive markers.

Acknowledgments

This work has been funded by the Deutsche Forschungsgemeinschaft (DFG) within the project "How to Win Arguments - Empowering Virtual Agents to Improve their Persuasiveness", Grant Number 376696351, as part of the Priority Program "Robust Argumentation Machines (RATIO)" (SPP-1999).

⁵ We plan to make the extended corpus along with the source code publicly available in the upcoming weeks.

References

1. Adadi, A., Berrada, M.: Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access* **6**, 52138–52160 (2018)
2. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity Checks for Saliency Maps. In: *Advances in Neural Information Processing Systems* 31, pp. 9505–9515. Curran Associates, Inc. (2018)
3. Alber, M., Lapuschkin, S., Seegerer, P., Hägele, M., Schütt, K.T., Montavon, G., Samek, W., Müller, K.R., Dähne, S., Kindermans, P.J.: investigate neural networks. *Journal of Machine Learning Research* **20**(93), 1–8 (2019)
4. Anders, C.J., Montavon, G., Samek, W., Müller, K.R.: Understanding patch-based learning of video data by explaining predictions. In: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pp. 297–309. Springer (2019)
5. Andrist, S., Spannan, E., Mutlu, B.: Rhetorical robots: making robots more effective speakers using linguistic cues of expertise. In: *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. pp. 341–348. IEEE (2013)
6. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLOS ONE* **10**(7) (Jul 2015)
7. Baur, T., Heimerl, A., Lingenfelser, F., Wagner, J., Valstar, M.F., Schuller, B., André, E.: explainable cooperative machine learning with nova. *KI-Künstliche Intelligenz* pp. 1–22 (2020)
8. Chaiken, S.: Heuristic and systematic information processing within and beyond the persuasion context. *Unintended thought* pp. 212–252 (1989)
9. Chidambaram, V., Chiang, Y.H., Mutlu, B.: Designing persuasive robots: how robots might persuade people using vocal and nonverbal cues. In: *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*. pp. 293–300 (2012)
10. DeSteno, D., Petty, R.E., Rucker, D.D., Wegener, D.T., Braverman, J.: Discrete emotions and persuasion: the role of emotion-induced expectancies. *Journal of personality and social psychology* **86**(1), 43 (2004)
11. Donadello, I., Dragoni, M., Eccher, C.: Persuasive explanation of reasoning inferences on dietary data. In: Demidova, E., Dietze, S., Breslin, J.G., Gottschalk, S., Cimiano, P., Ell, B., Lawrynowicz, A., Moss, L., Ngomo, A.N. (eds.) *Joint Proceedings of the 6th International Workshop on Dataset PROFILING and Search & the 1st Workshop on Semantic Explainability co-located with the 18th International Semantic Web Conference (ISWC 2019)*, Auckland, New Zealand, October 27, 2019. *CEUR Workshop Proceedings*, vol. 2465, pp. 46–61. CEUR-WS.org (2019)
12. Escalante, H.J., Guyon, I., Escalera, S., Jacques, J.C.S., Madadi, M., Baró, X., Ayache, S., Viegas, E., Güçlütürk, Y., Güçlü, U., van Gerven, M.A.J., van Lier, R.: Design of an explainable machine learning challenge for video interviews. In: *2017 International Joint Conference on Neural Networks, IJCNN 2017*, Anchorage, AK, USA, May 14–19, 2017. pp. 3688–3695. IEEE (2017)
13. Greydanus, S., Koul, A., Dodge, J., Fern, A.: Visualizing and understanding atari agents. In: *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden*. pp. 1787–1796 (2018)
14. Ham, J., Bokhorst, R., Cuijpers, R., van der Pol, D., Cabibihan, J.J.: Making robots persuasive: the influence of combining persuasive strategies (gazing and gestures) by a storytelling robot on its persuasive power. In: *International conference on social robotics*. pp. 71–83. Springer (2011)

15. Huber, T., Schiller, D., André, E.: Enhancing explainability of deep reinforcement learning through selective layer-wise relevance propagation. In: Joint German/Austrian Conference on Artificial Intelligence (Künstliche Intelligenz). pp. 188–202. Springer (2019)
16. Kaptein, M., Lacroix, J., Saini, P.: Individual differences in persuadability in the health promotion domain. In: International Conference on Persuasive Technology. pp. 94–105. Springer (2010)
17. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
18. van Kleef, G.: Emotions as agents of social influence. In: The Oxford Handbook of Social Influence. Oxford University Press (2019)
19. Kotikalapudi, R., contributors: keras-vis. <https://github.com/raghakot/keras-vis> (2017)
20. Krapinger, G.: Aristoteles: Rhetorik. Übersetzt und herausgegeben von Gernot Krapinger. Stuttgart: Reclam (1999)
21. Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., Müller, K.R.: Unmasking clever hans predictors and assessing what machines really learn. *Nature Communications* **10**(1), 1096 (2019)
22. Maricchiolo, F., Gnisci, A., Bonaiuto, M., Ficca, G.: Effects of different types of hand gestures in persuasive speech on receivers’ evaluations. *Language and cognitive processes* **24**(2), 239–266 (2009)
23. Molnar, C.: Interpretable Machine Learning. Lulu. com (2019)
24. Montavon, G., Samek, W., Müller, K.: Methods for interpreting and understanding deep neural networks. *Digital Signal Processing* **73**, 1–15 (2018)
25. O’Keefe, D.J., Jackson, S.: Argument quality and persuasive effects: A review of current approaches. In: *Argumentation and values: Proceedings of the ninth Alta conference on argumentation*. pp. 88–92. Speech Communication Association Annandale (1995)
26. Petty, R.E., Cacioppo, J.T.: The elaboration likelihood model of persuasion. In: *Communication and persuasion*, pp. 1–24. Springer (1986)
27. Poggi, I., Vincze, L.: Gesture, gaze and persuasive strategies in political discourse. In: *International LREC Workshop on Multimodal Corpora*. pp. 73–92. Springer (2008)
28. Ribeiro, M.T., Singh, S., Guestrin, C.: ”why should I trust you?”: Explaining the predictions of any classifier. In: Krishnapuram, B., Shah, M., Smola, A.J., Aggarwal, C.C., Shen, D., Rastogi, R. (eds.) *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, August 13–17, 2016. pp. 1135–1144. ACM (2016)
29. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: *Proceedings of the IEEE international conference on computer vision*. pp. 618–626 (2017)
30. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. *CoRR* **abs/1312.6034** (2013)
31. Sixt, L., Granz, M., Landgraf, T.: When explanations lie: Why modified BP attribution fails. *CoRR* **abs/1912.09818** (2019)
32. Van Kleef, G.A., van den Berg, H., Heerdink, M.W.: The persuasive power of emotions: Effects of emotional expressions on attitude formation and change. *Journal of Applied Psychology* **100**(4), 1124 (2015)

33. Wang, Y., Lucas, G., Khooshabeh, P., De Melo, C., Gratch, J.: Effects of emotional expressions on persuasion. *Social Influence* **10**(4), 236–249 (2015)
34. Weitz, K., Hassan, T., Schmid, U., Garbas, J.U.: Deep-learned faces of pain and emotions: Elucidating the differences of facial expressions with the help of explainable AI methods. *tm-Technisches Messen* **86**(7-8), 404–412 (2019)
35. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, Proceedings, Part I*. pp. 818–833 (2014)
36. Zhang, Y., Chen, X.: Explainable recommendation: A survey and new perspectives. *Found. Trends Inf. Retr.* **14**(1), 1–101 (2020)